DOCUMENT RESUME

ED 243 586 PS 014 279

AUTHOR Keesling, J. Ward; Smith, Allen G.
TITLE Issues Related to Instrumentation in Large-Scale Program Evaluation: Research Needs, Selection Considerations, and Alternative Outcome Indicators.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 80
NOTE 14p.; Filmed from best available copy. Paper presented at the Learning Research and Development Center/National Institute of Education Conference on Follow Through (Pittsburgh, PA, March 12-13, 1980).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Agency Role; Early Childhood Education; *Program Evaluation; *Research Design; *Research Methodology; *Research Needs; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS National Institute of Education; *Project Follow Through

ABSTRACT

Future research in Follow Through should be eclectic and comprehensive. In particular, the setting for Strand Two of the National Institute of Education's (NIE) research program for Follow Through should provide opportunities for sponsors of innovations to implement programs targeted at many outcomes. There should be a commitment to measuring varying outcomes in order to relate them to characteristics of a program or model and to the conditions of its implementation. For evaluation purposes, a common core of measures should be used; matrix sampling methods might be used to assess performance on those measures. Participation in the new research initiative will require the sponsor to name the outcomes to be influenced by the program. Further, the sponsor will be required to propose a measure or measures for assessing those outcomes. One of the services that the new Follow Through research program could perform would be to help sponsors create adequate alternative measures of outcomes. Follow Through program sponsors, rather than NIE alone, should be involved in the process of instrument selection and development. The standard terminology of validity and reliability is relevant to the test selection and development effort. (A list of some potential outcomes of Project Follow Through is appended.) (RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED243586

PS 014279

Issues Related To Instrumentation in Large-Scale Program Evaluation: Research Needs, Selection Considerations, and Alternative Outcome Indicators

J. Ward Keesling and Allen G. Smith
System Development Corporation

The goal of this paper is to describe, in a very general way, a setting within which we believe a wide variety of educational innovations may be tried out and evaluated. Within the context of this setting, we will discuss issues related to the selection and development of alternative outcome indicators.

In general, the position taken here will be that future research in Follow Through needs to be eclectic and comprehensive. We take this position, recognizing that such aspirations are by no means new to Follow Through. As Haney (1977) has said, the recent national evaluation of Follow Through set out with the best and broadest of intentions. Early plans called for evaluation of a broad range of parent and child outcomes, for a description of classroom processes and implementation, for an evaluation of institutional outcomes, and so on. In spite of these intentions, however, the scope of that study's analyses narrowed over time to a handful of child outcomes, assessed using only four measures. There were several reasons for this "funneling": conceptions of the goals for the program and for the evaluation were ambiguous and changing; measures of the alternative child outcomes and process variables were lacking and time was not available to develop new ones; models were newly developed, with objectives and treatments that were sometimes poorly defined. Although

the field of educational research has come a long way, any new evaluation must learn from Follow Through's past and be prepared to face similar pressures to narrow the scope of inquiry.

THE SETTING

In 1969 Donald Campbell described an "experimenting society" as one in which

> . . . we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available.

It is in this spirit that we describe our conceptualization of the setting for Strand Two of the NIE research program in Follow Through.

NIE has indicated that it wishes to encourage the development of innovative programs. It is important that a setting be provided that will accomodate a broad range of innovations so that attempts can be made to implement programs based on a wide variety of theoretical premises. It is also important that these programs be implemented in a variety of contexts so that information can be obtained about the ways in which the implementation of programs (and their outcomes) is conditioned by situation-specific factors.

As a part of this setting, NIE wishes to evaluate the implementations of these innovations to learn what features of them are related to various outcomes of interest. To some extent these evaluations will also provide information about the theories that form the basis for the innovations. For educational practice, it may be more important to assess the relationships of features of the implementations to the outcomes than it is to assess the theories themselves. NIE has taken

the position (that seems consonant with the data) that many different programs (called models in Follow Through) appear to be able to produce effects on certain outcome measures. Educational practitioners may benefit most by learning what forms of implementation are effective -- especially, what forms of implementation are best suited to particular contexts.

We believe that a setting which would encourage both innovation and appropriate evaluation can be developed. It requires a source of innovations and an acceptance of alternative outcomes (or alternative measures of traditional outcomes). Innovations can be supported by direct Federal sponsorship; imagine NIE developing such an innovation and implementing it in schools. Or, innovations can be supported by state, LEA or private sponsorship. The evidence of the Follow Through experience thus far is that many such agencies will volunteer to sponsor programs. In Strand Two of the new research program in Follow Through, a broad base of sponsorship should be encouraged in order to generate approaches to as many educational problems as possible.

Accompanying this broad base of sponsorship will have to be a willingness to accept a broad range of targeted outcomes, and a commitment to developing measures of those outcomes. This seems self-evident, for what purpose could be served by encouraging innovations with many targeted outcomes and then evaluating performance on only a very few? Yet, this is just what happened in the last evaluation of Follow Through.

In summary, the setting should provide opportunities for several sponsors of innovations to implement programs targeted at many outcomes. There should be a commitment to measuring these varying outcomes in order to relate them to the characteristics of the program (or model) and the conditions of its implementation. There are also conditions that should be placed on sponsors who wish to participate in such a research setting, and these will be discussed in a subsequent section.

THE CASE FOR A COMMON CORE OF MEASURES

The suggestion that many outcome indicators are desireable will cause many of us (including at least one of the authors of this paper) to feel the tension between the canon of parsimony and the call to proliferation of measures. We propose to relieve some of this tension by having a common core of measures. But first, we review some of the arguments for parsimony.

The first argument, as applied to proliferating measures of achievement, is that there are really only a very few dimensions of human intellectual performance. Therefore, any new measure that is proposed must demonstrate (via a step-wise testing procedure, say) that it contains information not found in measures of the known dimensions of importance. Neither the number of essential dimensions, nor their nature, is firmly established, however (polar extremes being represented by Spearman's G-Theory and Guilford's Structure of Intellect). Moreover, shifts in curricular emphasis may be more likely to show up as changes in patterns of performance on several measures of small segments of the entire curriculum.

A second argument is that in practice there are only a few outcomes of real importance. It is argued that the public and their representatives really only want to know what happened to the reading scores this year (a few may interest themselves in mathematics as well). Perhaps this should be seen as a result of our inexperience in valuing other outcomes, because we have had so little chance to deal with measures of them. This can be turned into a benefit of the research setting described above: A broader group of people will be sensitized to measures of alternative outcomes and have to deal with assessing their value vis-a-vis other measures.

A third, and related, argument that held sway in the earlier evaluation of Follow Through, is that there is a political need to have a "common thread" among all the models so that a horse race among sponsors can be

run, and judged. In the earlier evaluation this argument was used to justify nationally normed standardized tests as the common measure because they were understood by the public and their representatives. Other papers in this session discuss the problems and prospects of such tests, so we will not dwell on this issue.

The argument that leads us to propose a common core of measures is that there are some features of the curriculum of the primary grades that simply cannot be ignored. If they were ignored, potential beneficial or detrimental effects of particular importance might be missed. To be more concrete, the end of the primary grades is when reading skills are generally becoming well developed; it would remiss not to measure reading skills at the end of grade three. The shifts in curricular emphasis (potentially combined with new forms of instruction) could yield benefits or deficits that would be missed if the instruments used to evaluate the model focused on mathematics and human relations.

In proposing a common core, we are not proposing the specific content of that core. We believe that a consensual process should be used to determine what the content should be. This might be accomplished by having the current sponsors nominate and (it is hoped) agree upon some small set of outcomes to be measured. They might be encouraged to do this if it is clear that they may choose additional goals to be used in the assessment of their own programs.

Of course, even if some agreement can be reached on the outcomes of interest, this does not guarantee agreement on the instruments to be used to measure the outcomes. In order to keep testing time down, however, some concensus should be forged at the outset. We will indicate a way in which new measures can be developed and assessed in a subsequent part of this paper.

Matrix sampling methods might be used to reduce the time required to assess performance on the common core measures. The success of the California Assessment Program suggests that this is a viable

alternative, if one is willing to give up certain other uses of the measures (such as diagnostic information for the individual students).

The common core might also include some process measures. For example, if certain achievement outcomes are regarded as valuable enough to measure, it might be worthwhile to assess the amount of instructional time offered in these subjects. One could go even farther and inquire about the time attended (and reasons for non-attendance), and the amount of active learning time. Differences in outcomes across programs could be explained by differences in these time measures. The time measures themselves might be related to certain features of the program implementation or of the environments in which it is implemented.

Of course, there are risks to having a common core of measures. The foremost is that this core will be seen to contain the essence of the goals of the Follow Through program, and the results from the other measures will be ignored. This will result in the same kind of "funneling-down" that occurred in the previous evaluation of Follow Through. A related problem is that it may be difficult to develop measures of new outcomes that will appear to be of high enough quality to merit attention. This is the problem that plagued the earlier evaluation effort: it became too expensive to develop and use the new measures, and, over time, the set of measures funneled-down to a few well-established paper-and-pencil instruments.

The only antidote we can see to these risks is that NIE make a visible commitment to the development of alternative measures of core outcomes, as well as to measures of other outcomes. It must be clear from the start that the alternative measures will be accorded the time and resources needed to develop instruments of high quality.

THE SPECIFICATION OF ALTERNATIVE OUTCOMES

Implicit in the earlier sections of this paper was the notion that sponsors should specify the outcomes that they believe their models will affect. Thus, participation in the new research initiative in Follow Through will require more than passive acceptance of the common core of measures; it will require the sponsor to name the outcomes to be influenced by the program. A further requirement, to be discussed later, is that the sponsor propose a measure (or measures) of those outcomes. This contrasts dramatically with the situation in the last evaluation of Follow Through where many sponsors did not worry about defininng goals, much less measures of the outcomes.

The concensus building process suggested as a means to selecting the common core of outcomes and measures could also be used to develop a pool of alternative outcomes and measures. It is our idea that this pool would, over time, be a continuously self-augmenting "cafeteria" from which to make choices. Each new sponsor would be responsible for a truly "planned" variation, complete with goal specification and proposed method for assessing outcomes.

Of course, totally laissez-faire development of new programs may not be the most efficient way to conduct this research initiative. NIE (perhaps with the assistance of current sponsors) could develop a review process that would screen proposals for their relevance to particular concerns and their uniqueness compared to the other models in the experiment. NIE could be ever more directive by requesting the development of models with particular emphases. In this way NIE could assure that important lines of inquiry will be addressed.

Some Possible Alternative Goals

Follow Through provides funds to support "comprehensive educational, helath, nutritional, social and other services as will aid in the continued development of children (from low-income families). . . to their full potential." It is desireable to indicate some of the goals

that might be specified as outcomes of models to give a sense of the scope of possible interventions. We don't wish to propose any particular value ordering in presenting this list. We feel that such valuing should emerge in a dynamic process involving NIE, the research community, and potential sponsors.

Table 1, adapted from Hodges, Branden, Feldman, Follins, Love, Sheehan, Lumbley, Osborn, Rentfrow, Houston and Lee (1980), should be considered only a partial listing of outcomes of interest. (Research conducted by the Follow Through sponsors themselves on alternative outcomes and measures covers about 5 linear feet of shelf space.) Many other aspects of programs and their implementation could be studied in order to understand the processes that influence the outcomes we have tabled.

To repeat a previous point, some process measures, such as those listed under Service Delivery Outcomes should be measured in order to understand how the programs affect student outcomes. This information will be especially valuable when comparing models to one another.

The publication of Lasting Effects After Preschool (Lazar and Darlington, 1979) reminds us that many outcomes of Follow Through should be assessed longitudinally -- including years beyond the third grade. As a part of its research mission, NIE should consider supporting further research into the long term effects of these early interventions.

Instrumentation for Measures of Alternative Outcomes

Demanding that sponsors indicate the outcomes they expect their models to influence immediately raises the issue of what measures of those outcomes should be used. One aspect of this question is the quality of the instruments, which we will say more about shortly. Another aspect is the acceptability of the measures to the sponsors.

Some measures are likely to be a closer match to the content of a particular sponsor's curriculum, be less expensive to administer, and use formats for presenting problems and recording responses that are more consistent with the common experiences of the children being served. Any or all of these reasons could be used to justify the selection of one measure and the rejection of others.

One of the services that the new research program in Follow Through could perform would be to aide in the creation of alternative measures of outcomes. We believe that the initiative should come from the sponsors themselves. They should propose measures for assessing the outcomes they specify as targets of their programs. If the quality of these instruments for use with the subject population of Follow Through is not well known, then the sponsor and NIE should jointly engage in an evaluation of their quality. NIE could provide technical assistance on measurement to the sponsors, and could serve as a broker in linking sponsors who would be willing to help one-another by providing additional subjects and settings for trials of the instruments.

NIE should carefully consider the value of the proposed instruments to all potential users before agreeing to co-sponsor their development. Is there a unique feature that would make it more suitable than other instruments already in use? Is it more cost-effective? Would it be useful in other programs? Ultimately there might be a review process similar to the Joint Dissemination Review Panel (JDRP) that would pass on the acceptability of the instrument for general use in Follow Through research.

One way to manage this effort would be to contract with one group, who would be experts in measurement, to work with sponsors on the development of new instruments. Only sponsors would be permitted to apply for and receive instrument development contracts, which would flow through the managing group.

There are evident risks in this method for developing instruments. Not the least of which is the time that will be needed to try out and review the results of various instruments. The "bootstrapping" nature of the process also creates risks. A sponsor will be simultaneously developing an instrument and implementing a program to affect the outcomes measured by that instrument. If the instrument shows no changes in time or no differences across groups, it may be difficult to determine whether the instrument is defective or the program was not well implemented, or was ineffective.

However, we feel that it is preferable to involve sponsors in this process, rather than to have NIE alone select or develop instruments. If the sponsors have some ownership of the measures, they may be more willing to treat the results of their use more candidly.

ASSESSING THE QUALITY OF INSTRUMENTS

A great deal of research has been performed lately on the proper way to assess the quality of measures of academic aptitude and achievement. Starting from the point of view of domain-referenced tests or from the point of view of item characteristic curve theory, elaborate structures can be generated to assess the quality of test-like instruments. Other papers in this series deal with the development of tests, and we wish to make comments that generalize to a broader collection of instruments, so we will not provide further details on these methodologies.

We will, instead, return to the standard terminology of validity and reliability and give capsule statements of their meaning in the context of the new research program in Follow Through.

Validity

Traditionally, one thinks of four types of validity:

Content Validity deals with the extent that the measure represents the content of the universe to which one wishes to generalize. A measure of total caloric intake should attend to all of the sources of nutrients to which a child is exposed, for example. One criticism of the earlier evaluation of Follow Through was that the content of the standardized tests was not an appropriate sample of the behaviors that the Follow Through models were attempting to alter. The goal was not to improve performance on the seemingly artificial paper-and-pencil, multiple choice tests, but to improve performance in more realistic endeavors such as communicating verbally to one another and to the teacher.

Concurrent Validity deals with the extent to which a measure represents the same universe as another measure. This will be a critical issue in the new research program, because many measures may be proposed for the same set of content. It will be important to allow some duplication of content to permit the construction of cost-effective alternatives and to provide model sponsors opportunities to cross-check or triangulate their results with alternative measures. Supporting many alternative measures of the same content may be counter-productive, however. It would be better to expand the scope of evaluation to cover important programatic emphases thoroughly, rather than to replicate measures in established areas. There will be some difficult decisions to make along these lines.

Predictive Validity deals with the extent that performance on an instrument given in the present will improve the precision of a guess about performance in the future. This is related to the idea of longitudinal assessment and should be a research thrust in the new NIE program. If the measures of outcomes used during the Follow Through years are not related to measures of other outcomes used in later years,

then what have we demonstrated by showing that programs can influence the earlier measures?

Construct Validity deals with the extent to which the measure responds to variations in variables that would be expected to influence the underlying construct. Are the measures of achievement outcomes sensitive to instruction? This recalls the problem of the bootstrapping nature of the research program described in this paper. This is the persisting paradox of construct validity. How to be sure that the treatment was effective when the measure shows no response?

Reliability

The primary issue we wish to raise here is the notion of the level at which a reliable measurement is desired. Some measures proposed in Table 1 will be taken at the classroom or school level, by necessity. Is it important to have all of the measures of child-associated outcomes be accurate at the level of the individual child? Earlier we indicated that a matrix-sampling approach could be used to reduce the testing burden. Of course, this will prevent certain types of analysis from being done. For example, assessing the progress of individual students longitudinally will not be possible.

In general, we feel that generalizability theory provides the answers to the questions about reliability, if it is possible to determine ahead of time what the decision-studies will be that will make use of the instruments under development. If the nature of the decision studies is not specifiable, then it will be difficult to know whether the generalizbility studies are more than an exercise for the numerically sophisticated.

Table 1. Some Potential Outcomes of Follow Through

Child Outcomes

Achievement (reading, math, language)

Personal Development (self concept, locus of control, initiative, independence) attitudes toward school)

Health and Fitness

Need for special educative services (placement in special education, retention in grade)

Parent Personal Outcomes

Parenting Skills

Teaching Skills

Decision-making Skills

Socio-economic Advancement

Attitudes toward Schools

Paraprofessional Personal Outcomes

Teaching Skills

Career Advancement

Socio-economic Advancement

Attitudes toward School

Professional Personal Outcomes

Teaching Skills

Administrative Skills

Attitudes toward project children and their families

Institutional Outcomes

Relationships among Ed. Agencies

Curricular Changes

Classroom Management Practices

Parent Involvement

Information Utilization

Service Delivery Outcomes

Time engaged in on-task behaviors

Patterns of ed. services offered

Ancillary (e.g., health) services offered